1Б

УДК 519.237

Б.С. Бусыгин, Е.П. Зацепин
Национальный горный университет, г. Днепропетровск, Украина
busyginb@yandex.ru

Метод трехмерной триангуляции в задачах кластерного анализа

Представлен метод многомерного шкалирования на основе трехмерной триангуляции. Рассмотрена возможность качественного сохранения геометрической структуры множества объектов при отображении многомерного пространства в трехмерное. Приведены результаты применения метода для решения задачи кластеризации на примере периодической системы элементов Д.И. Менделеева. Экспериментально показано, что по критериям качества кластеризации предложенный метод более эффективен в сравнении с методами k-средних и нейронной сети Кохонена.

Введение

В настоящее время для решения задач анализа данных применяется широкий набор методов кластерного анализа, среди которых наибольшее распространение получили методы сортировки, оптимизации критериев качества, иерархические (агломеративные и дивизимные), неиерархические (итеративные) методы, а также методы многоэтапной кластеризации [1-3]. Большинство перечисленных методов характеризуются сложностью вычислительной реализации, а результаты их применения, как правило, зависят от структуры данных в многомерном признаковом пространстве, что обусловливает нецелесообразность их применения при выявлении кластеров сложной формы. Кроме того, к недостаткам некоторых распространенных методов также относятся необходимость задания пороговых значений и количества кластеров, чувствительность к выбросам.

Указанные обстоятельства привели к необходимости поиска таких методов кластеризации, которые отличались бы максимальной независимостью результатов от структуры исходных данных в признаковом пространстве. Один из них базируется на основе нейронной сети Кохонена [4], которая, однако, характеризуется существенными вычислительными затратами, обусловленными используемой итеративной процедурой самоорганизации. Другим популярным методом кластеризации является метод нечетких С-средних [5], позволяющий с большой точностью классифицировать объекты, лежащие на границах выделяемых кластеров, хотя и этот метод также не лишен недостатков: помимо необходимости задания числа кластеров, зачастую возникает неопределенность с объектами, удаленными от центров всех кластеров.

В качестве альтернативы методам кластеризации в задачах анализа данных могут использоваться методы многомерного шкалирования [6-9], заключающиеся в понижении размерности признакового пространства с сохранением геометрической структуры исходного множества. Эффект достигается за счет применения процедуры элементарных геометрических построений, в ходе которых все точки исходного множества отображаются в пространстве малой размерности ($K \leq 3$) с целью достижения аппроксимации исходной матрицы расстояний (различий) между объектами в многомерном признаковом пространстве. Такое отображение дает возможность представления результатов кластеризации в удобной для визуального анализа форме.

Цель настоящей работы – описание и демонстрация возможностей метода трехмерной триангуляции на примере кластеризации элементов периодической системы Д.И. Менделеева. Указанный метод является модификацией двухмерного варианта триангуляции, представленного в работе [10].

Описание метода

Пусть даны четыре точки m_1, m_2, m_3 и m_4 многомерного пространства. Точки m'_1, m'_2, m'_3 в трехмерном пространстве выбираются таким образом, что разделяющие их расстояния равны соответствующим попарным расстояниям между точками m_1, m_2, m_3 в многомерном пространстве. Четвертая точка m_4 многомерного пространства отображается в точку m'_4 в трехмерном пространстве с таким расчетом, чтобы расстояние между m_4 и точками m_1, m_2, m_3 сохранялось при отображении неизменным. Для ее нахождения выполняется поиск точки пересечения трех сфер с радиусами d_{14}, d_{24}, d_{34}, где d_{ij} – расстояние между точками m_i и m_j многомерного пространства. На рис.1 показано, как находится точка m'_4. Следует отметить, что если точки m_1 – m_4 не лежат в одной плоскости, то m_4 соответствуют две точки m'_4.

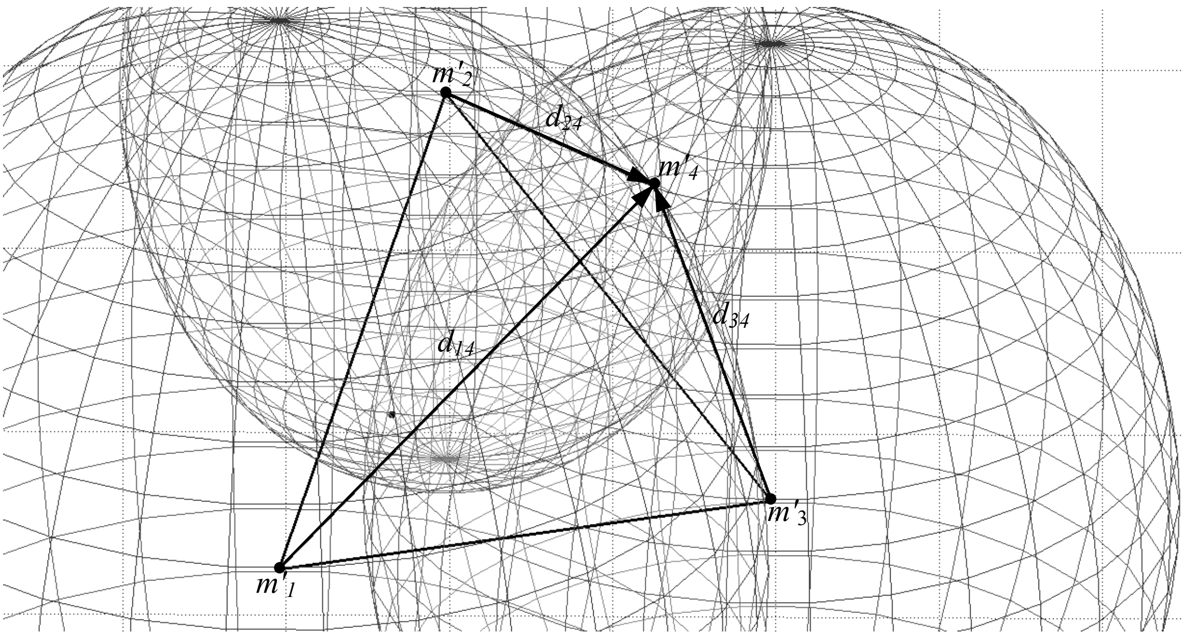


Рисунок 1 – Отображение четырех точек многомерного пространства в трехмерное путем нахождения пересечения трех сфер

Может возникнуть ситуация, когда расположение точек в пространстве не удовлетворяет неравенствам треугольника. Для устранения такого рода проблем используется метод аддитивной константы [11], основанный на предположении, что если ко всем расстояниям между точками прибавить некоторую константу, то они в равной степени будут отражать пространственную структуру объектов.

Далее путем последовательного построения выполняется отображение множества N точек многомерного пространства в трехмерное. Для каждой новой отображаемой точки (кроме трех первых) имеется три расстояния, которые необходимо выдержать. Правильный выбор этих расстояний позволяет сохранить существенную часть информации о геометрических взаимоотношениях между отображаемыми точками в трехмерном представлении. Для этой цели подходящими объектами служат $N – 1$ ребер

минимального остовного дерева, построенного для данного множества точек, поскольку минимальное остовное дерево несет в себе наиболее существенную часть структурной информации, заложенной в данных, и представляет её в компактной форме.

В рамках метода триангуляции используется два подхода, определяющих, какие из опорных ребер следует сохранить.

Триангуляция на основе второго ближайшего соседа. При выполнении отображения точки m_4 сохраняются расстояния d_{14}, d_{24}, d_{34}, где $m_1 - m_3$ – точки, ранее отображенные в пространство и удовлетворяющие следующим условиям: а) в минимальном остовном дереве точка m_3 непосредственно соединена с точкой m_4; б) m_1, m_2 – те соседние точки для m_3, которые в минимальном остовном дереве ближе других соседних точек расположены к точке m_4.

Триангуляция на основе точки отсчёта. Тремя расстояниями, сохраняемыми при отображении точки, являются: длина двух ребер минимального остовного дерева и расстояние до произвольной, заранее выбранной точки отсчета. Это позволяет получить N трехмерных представлений данных, тем самым обеспечивается возможность исчерпывающего анализа кластеризации отображаемых объектов.

Экспериментальные результаты

Для демонстрации возможностей метода триангуляции использовалась периодическая система элементов Д.И. Менделеева, которая характеризуется высокой степенью разнородности объектов. В качестве исследуемого множества взяты 75 химических элементов, а в качестве признаков – 14 свойств двух основных типов [12]: атомные (атомная масса, атомный объём, атомный радиус, ионный радиус, период решетки, плотность, первый потенциал ионизации, электроотрицательность) и термодинамические (температура плавления, температура кипения, характеристическая температура Дебая, удельная теплота плавления, удельная теплота испарения, удельная теплоёмкость).

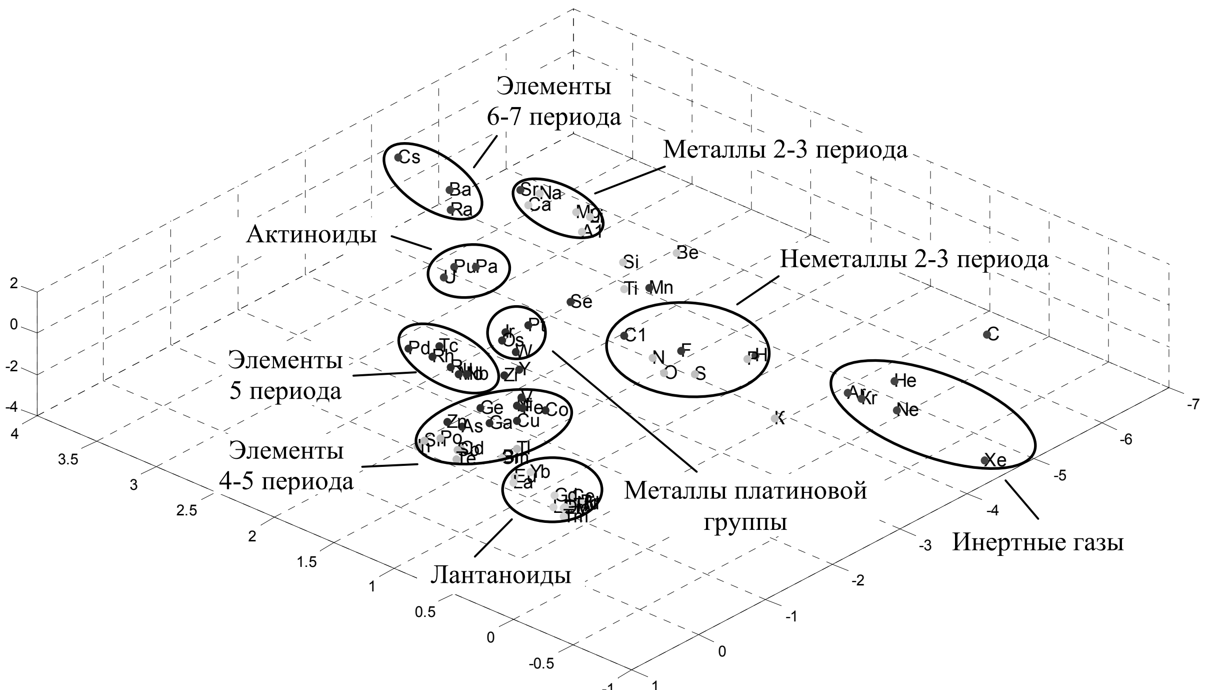


Рисунок 2 – Отображение объектов многомерного пространства в трехмерное (триангуляция на основе второго ближайшего соседа)

На рис. 2, 3 представлено отображение объектов (химических элементов) в трехмерное пространство, полученное в рамках применения метода триангуляции на основе второго ближайшего соседа с применением Евклидовой меры расстояния. Необходимо отметить, что оси на получаемых отображениях играют роль показателя уровня кластеризации и позволяют осуществить визуальную оценку различий между объектами. Началом координат служит точка отсчета (при использовании одноименного подхода); при использовании триангуляции на основе второго ближайшего соседа в начале координат располагается точка с минимальной суммой расстояний до двух ближайших точек.

На рисунках хорошо видны кластеры неметаллов 2 – 3 периода (N, O, F, Cl, S, P, H) с отделением инертных газов (He, Ne, Ar, Kr, Xe). Кроме того, отчетливо выделяются металлы платиновой группы (Os, Ir, W, Pt), заметно отличающиеся показателями температур плавления, кипения и плотности. Просматривается кластеризация периодической системы по периодам.

Отображения, получаемые триангуляцией на основе точки отсчета (рис. 4 – 6), дают возможность проанализировать пространственное расположение объектов «со всех сторон», позволяя выделить незаметные объединения, а также судить о взаимной удаленности точек друг от друга.

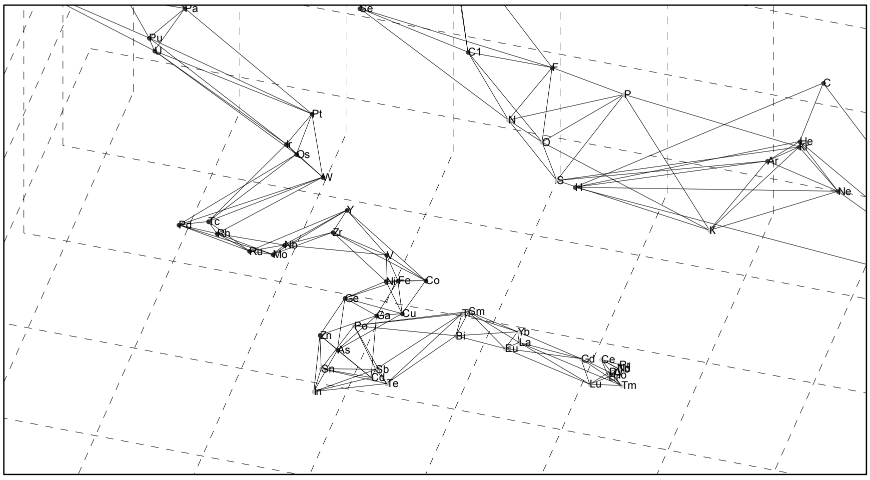


Рисунок 3 – Увеличенный фрагмент отображения с прорисовкой «опорных» треугольников

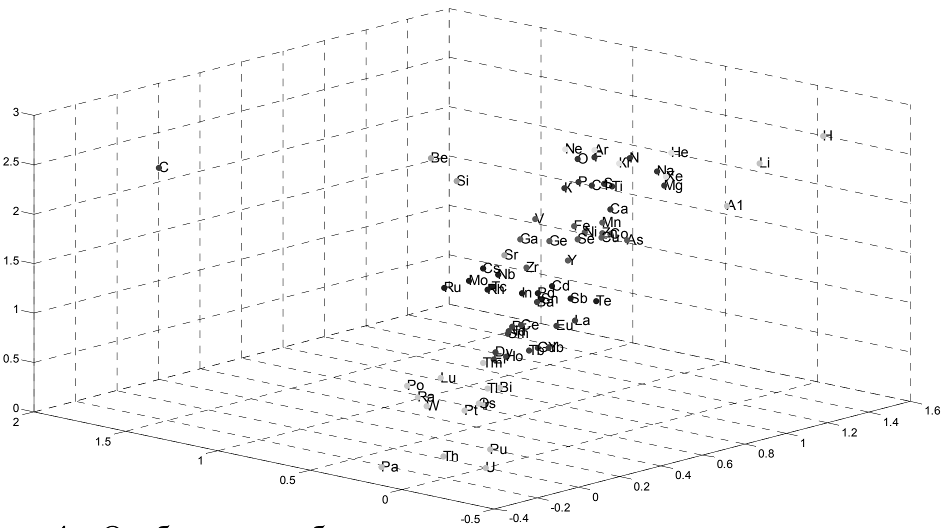


Рисунок 4 – Отображение объектов многомерного пространства в трехмерное (триангуляция на основе точки отсчета) с использованием объекта U (уран) в качестве точки отсчета

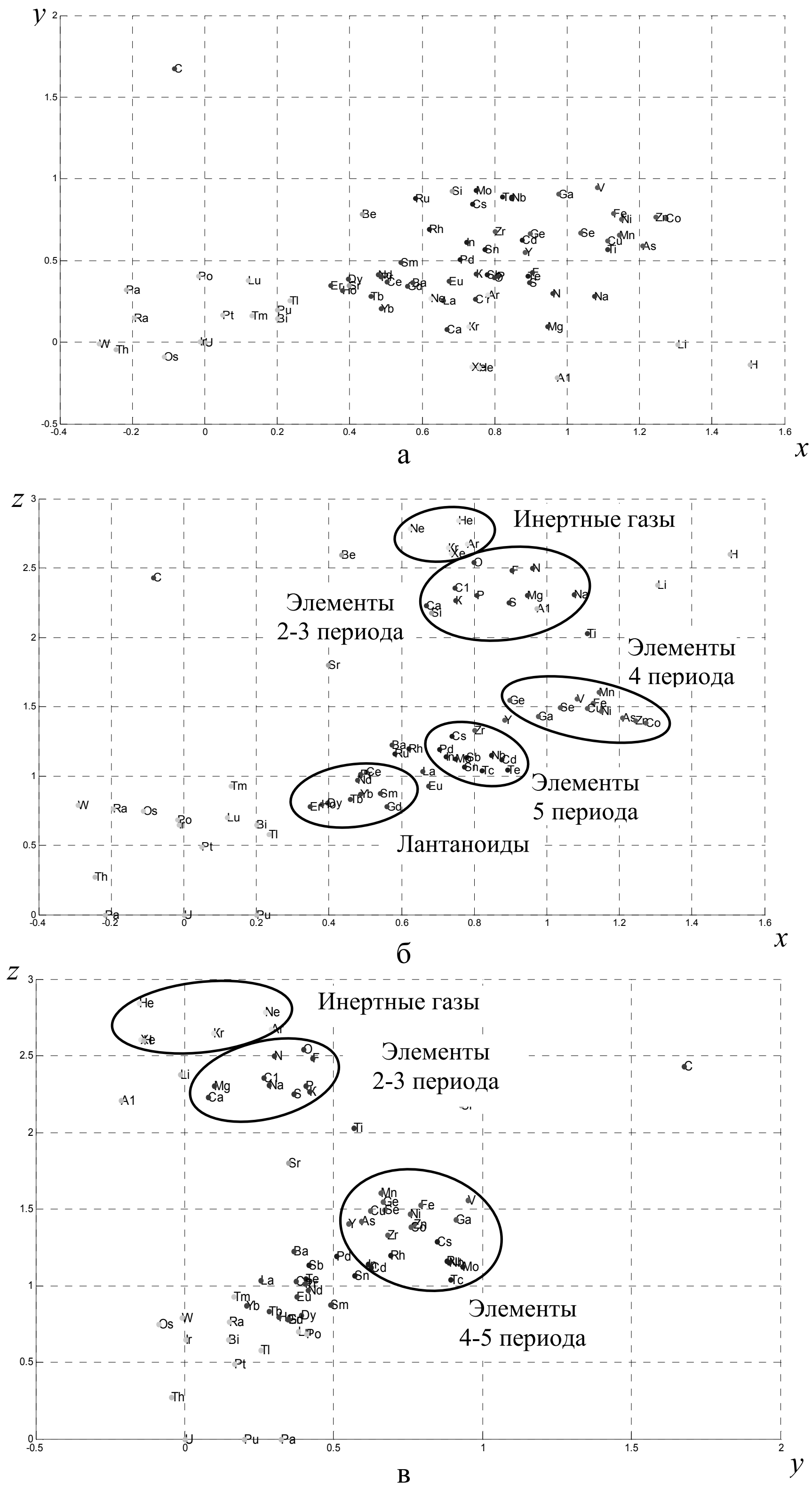


Рисунок 5 – Проекции отображения, представленного на рис. 4:
а – проекция *xy*, б – проекция *xz*, в – проекция *yz*

На рис. 4 – 6 видно, что общая структура получаемого отображения практически не изменяется по сравнению с отображениями, получаемыми триангуляцией на основе второго ближайшего соседа. По-прежнему наблюдается выделение наиболее отдален-

ных от точки отсчета кластеров (это хорошо заметно на проекциях отображения), что еще раз подчеркивает преимущества предложенного метода.

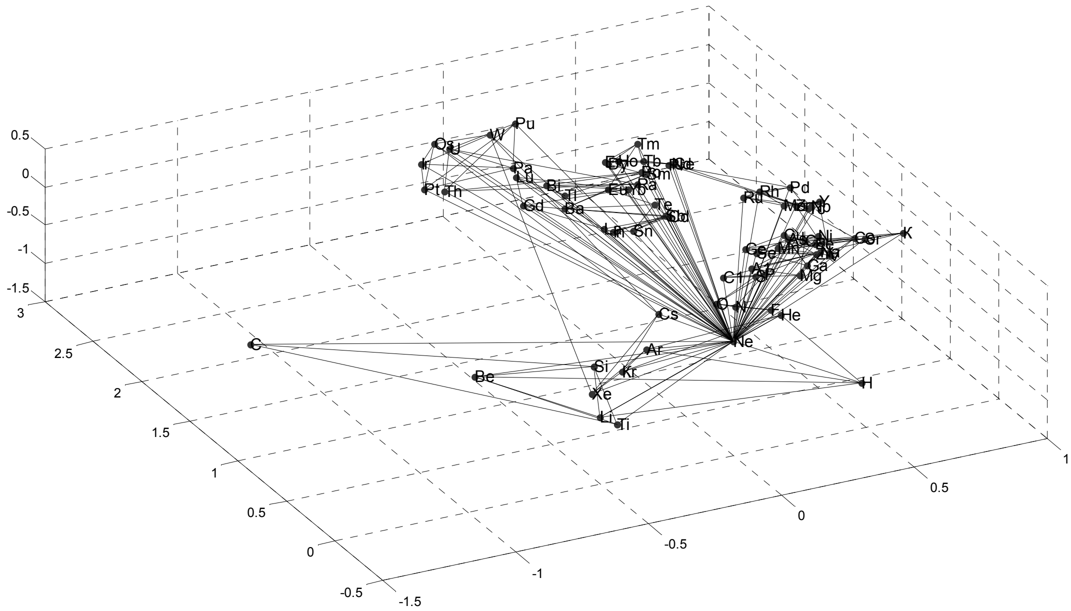


Рисунок 6 – Фрагмент отображения объектов многомерного пространства в трехмерное (триангуляция на основе точки отсчета) с использованием объекта Ne (неон) в качестве точки отсчета и отображением «опорных» треугольников

В целом применение метода трехмерной триангуляции позволило выделить следующие кластеры периодической системы элементов Д.И. Менделеева: неметаллы 2 – 3 периода (N, O, F, P, Cl); инертные газы (He, Ne, Ar, Kr, Xe); водород (H); металлы 2 – 4 периода (Li, K, Ca, Na, Mg, Al); осмий, иридий, вольфрам, платина (Os, Ir, W, Pt) – семейство платиновых металлов с очень высокими температурами плавления и кипения; элементы семейства железа (Fe, Co, Ni и пр.); элементы 5 периода, оксиды и гидроксиды которых проявляют амфотерные свойства (Y, Zr, Nb, Mo, Tc, Ru, Rh, Pd); элементы 6 – 7 периода за исключением платиновых металлов (Cs, Ba, Tl, Bi, Po, Ra); лантаноиды (La, Ce, Pr, Nd, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu); актиноиды (Th, Pa, U, Pu).

Для оценки качества полученной кластеризации использовались 22 показателя [13]: f_1 – сумма квадратов внутрикластерных расстояний; f_2 – среднее значение квадратов внутрикластерных расстояний; f_3 – сумма внутрикластерных расстояний; f_4 – среднее внутрикластерное расстояние; f_5 – сумма диаметров кластеров (диаметр кластера – максимальное внутрикластерное расстояние); f_6 – максимальный диаметр кластера; f_7 – сумма усредненных диаметров кластеров; f_8 – максимальный усредненный диаметр кластера; f_9 – сумма квадратов расстояний до центроида; f_{10} – среднее значение квадратов расстояний до центроида; f_{11} – сумма расстояний до центроида; f_{12} – среднее расстояние до центроида; f_{13} – сумма центроидных радиусов (центроидный радиус – наибольшее расстояние от центроида до объекта внутри кластера); f_{14} – максимальный центроидный радиус; f_{15} – сумма усредненных центроидных радиусов; f_{16} – максимальный усредненный центроидный радиус; f_{17} – сумма квадратов расстояний до медоида (медоид – объект кластера с минимальным средним расстоянием до других объектов внутри кластера); f_{18} – среднее значение квадратов расстояний до медоида; f_{19} – сумма минимальных расстояний между кластерами; f_{20} – сумма средних расстояний между кластерами; f_{21} – сумма расстояний между центроидами; f_{22} – сумма расстояний между медоидами.

Необходимо отметить, что меньшие значения показателей качества $f_1 - f_{18}$ свидетельствуют о большей компактности кластеров и, соответственно, о лучшем качестве кластеризации. Для показателей $f_{19} - f_{22}$, наоборот, качество кластеризации определяется максимумом значений критериев.

Полученные показатели качества сверялись с показателями качества кластеризации, произведенной для того же набора данных методами двухмерной триангуляции, нейронной сетью Кохонена и k-средних. Разбиение объектов на заданное число кластеров в алгоритмах двухмерной и трехмерной триангуляции выполнялось по методу Уорда [14] в двух- и трехмерном пространстве соответственно.

Таблица 1 – Показатели качества кластеризации для числа кластеров K = 5 и K = 15 (лучшие результаты по каждому критерию выделены подчеркиванием)

Критерий качества кластеризации	K = 5				K = 15			
	Триангуляция 2-D	Триангуляция 3-D	Алгоритм k-средних	Нейронная сеть Кохонена	Триангуляция 2-D	Триангуляция 3-D	Алгоритм k-средних	Нейронная сеть Кохонена
f_1	315,736	309,226	351,338	311,223	47,563	44,740	41,110	51,634
f_2	17,452	15,307	17,868	19,417	10,012	9,329	9,501	9,487
f_3	472,658	435,531	425,505	381,596	81,369	81,263	81,226	85,148
f_4	21,051	20,811	22,929	23,103	13,623	13,593	13,110	14,457
f_5	5,957	5,257	4,918	6,262	10,008	7,965	9,639	11,231
f_6	2,245	1,512	1,605	1,617	1,701	1,527	1,692	1,781
f_7	0,844	0,568	0,461	0,548	2,586	2,158	2,107	2,534
f_8	0,281	0,253	0,125	0,231	0,850	0,345	0,425	0,425
f_9	17,452	17,307	17,868	19,417	5,012	4,329	7,501	9,487
f_{10}	1,453	1,391	1,578	1,302	1,534	1,795	1,574	1,957
f_{11}	33,261	30,400	34,250	34,069	21,835	20,109	20,782	22,910
f_{12}	2,181	2,115	2,617	2,284	5,249	4,183	4,095	4,084
f_{13}	4,069	4,185	4,407	4,153	5,800	4,794	5,690	6,905
f_{14}	1,172	1,072	1,006	1,291	0,850	0,816	0,819	1,205
f_{15}	0,340	0,195	0,235	0,297	1,718	1,172	1,071	1,520
f_{16}	0,209	0,183	0,252	0,199	0,225	0,198	0,252	0,185
f_{17}	23,922	22,460	23,707	24,639	12,227	11,716	10,164	12,829
f_{18}	1,405	1,270	1,283	1,673	4,110	2,537	2,178	2,741
f_{19}	0,385	0,446	0,341	0,239	21,435	21,982	21,922	21,487
f_{20}	4,648	4,149	3,546	3,313	42,325	41,225	39,784	39,137
f_{21}	10,071	10,175	10,168	8,614	109,896	113,670	109,255	96,543
f_{22}	10,672	12,021	12,875	9,524	115,568	120,132	120,061	103,823

Соответствующие значения критериев (табл. 1) дают основание утверждать, что результаты кластеризации, полученные с применением метода трехмерной триангуляции, по большинству показателей качества превосходят результаты, полученные методами k-средних и нейронной сети Кохонена.

Выводы

1. Предложенный метод дает возможность качественного сохранения геометрической структуры множества объектов при сокращении размерности исследуемого пространства и позволяет уменьшить временные и вычислительные затраты на выполнение процедуры кластеризации.
2. Основным преимуществом метода является использование элементарных геометрических построений, а также отсутствие необходимости постоянного пересчета отображения.
3. Результаты оценки качества кластеризации свидетельствуют о незначительных потерях при сохранении кластерной структуры набора данных.
4. Метод трехмерной триангуляции позволяет визуализировать результаты кластеризации в 3D-пространстве, что дает возможность более содержательного анализа многомерных данных с использованием знаний и интуиции исследователя.
5. Выделяемые с помощью метода трехмерной триангуляции наборы кластеров периодической системы элементов соответствуют общепринятой классификации в соответствии с периодическим законом, устанавливающим закономерное изменение физических и химических свойств элементов по мере увеличения их атомной массы.
6. Для наиболее полного анализа, позволяющего свести к минимуму вероятность ошибочной кластеризации, следует последовательно анализировать отображения, где каждый объект выступает в роли точки отсчета с использованием одноименного подхода.
7. Триангуляция на основе второго ближайшего соседа может применяться в условиях, когда необходим быстрый анализ данных. Основное преимущество данного подхода – отсутствие затрат времени на полный перебор.
8. Полученный в результате применения метода триангуляции набор трехмерных данных может подвергаться дальнейшей обработке (например, для решения задач распознавания).

Литература

1. Jain A.K. Data Clustering: A Review / Jain A.K., Murty M.N., Flynn P.J. // ACM Computing Surveys. – 1999. – Vol. 31, № 3. – P. 264-323.
2. Жамбю М. Иерархический кластер-анализ и соответствия / Жамбю М. – М. : Финансы и статистика, 1988. – 342 с.
3. Romesburg H.C. Cluster Analysis for Researchers / Romesburg H.C. – Lulu Press, 2004. – 334 p.
4. Хайкин С. Нейронные сети: полный курс; [пер. с англ.] / Хайкин С. – 2-е изд. – М. : Издательский дом «Вильямс», 2006. – 1104 с.
5. Abonyi J. Cluster Analysis for Data Mining and System Identification / Abonyi J., Feil B. – A Birkhäuser book, 2007. – 303 p.
6. Терехина А.Ю. Анализ данных методами многомерного шкалирования / Терехина А.Ю. – М. : Наука. Главная редакция физико-математической литературы, 1986. – 168 с.
7. Терехина А.Ю. Неметрическое многомерное шкалирование. Препринт / Терехина А.Ю. – М. : ИПУ, 1977. – 73 с.
8. Misra H. Clustering Algorithm / Misra H. – TIFR, Mumbai, 1996. – P. 1-20.
9. Naud A. Interactive data exploration using MDS mapping / Naud A., Duch W. // Fifth Conference Neural Networks and Soft Computing. – Zakopane, Poland, 2000. – P. 255-260.
10. Бусыгин Б.С. Кластеризация объектов на основе применения метода триангуляции / Бусыгин Б.С., Зацепин Е.П. // Науковий вісник Національного гірничого університету. – 2006. – № 3. – С. 70-75.

11. Торгерсон У.С. Многомерное шкалирование. Теория и метод. Статистическое измерение качественных характеристик / Торгерсон У.С. – М. : Статистика, 1972. – С. 95-118.
12. Дриц М.Е. Свойства элементов : Справочник : в 2 т. / Дриц М.Е., Кузнецов Н.Т. – М. : Металлургия, 1997. – С. 310-332.
13. Nguyen Q.H. Internal quality measures for clustering in metric spaces / Nguyen Q.H., Rayward-Smith V.J. – Int. J. Business Intelligence and Data Mining. – 2008. – Vol. 3, № 1. – P. 4-29.
14. Ward J.H. Hierachical grouping to optimize an objective function / Ward J.H. – J. Am. Statist., 1963. – Assoc. 58. – P. 236-244.

Бусигін Б.С., Зацепін Є.П.

Метод тривимірної тріангуляції в задачах кластерного аналізу

Представлено метод багатовимірного шкалювання на основі тривимірної тріангуляції. Розглянуто можливість якісного збереження геометричної структури множини об’єктів при відображенні багатовимірного простору в тривимірне. Наведено результати застосування методу для вирішення задачі кластеризації на прикладі періодичної системи елементів Д.І. Менделєєва. Експериментально показано, що за критеріями якості кластеризації запропонований метод більш ефективний у порівнянні з методами k-середніх та нейронної мережі Кохонена.

B.S. Busygin, E.P. Zatsepin

The 3-D Triangulation Method in Problems of Cluster Analysis

The method of multidimensional scaling on the basis of the 3-D triangulation is presented. The qualitative preservation possibility of geometrical structure of objects by multidimensional space mapping to three-dimensional space is considered. The results of application of the method for clustering problem of the Mendeleyev periodic table are presented. It is experimentally shown, that the presented method is more effective by criteria of clustering quality in comparison with the methods of k-averages and the Kohonen neural network.

Статья поступила в редакцию 14.10.2009.